DOCUMENT RESUME

ED 261 085 TM 850 505

AUTHOR Medley, Donald M.
TITLE Issues and Problems in the Validation of Teaching and Teacher Professional Behaviors.
PUB DATE 3 Apr 85
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31- April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Gains; Classroom Research; Competency Based Teacher Education; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; Outcomes of Education; Prediction; Pretests Posttests; Reliability; Research Needs; Research Problems; *Teacher Behavior; *Teacher Characteristics; *Teacher Effectiveness; *Teacher Evaluation; Validity
IDENTIFIERS Gain Scores

ABSTRACT

There are a number of issues involved in the validation of measures of teacher effectiveness, that is, a teacher's ability to promote student learning. Teachers' characeristics may be assessed at five different points: (1) pre-existing teacher characteristics--knowledge, skills, and attitudes which are necessary for success in the teacher education curriculum and as a teacher; (2) teacher competence; (3) teacher performance within a specific classroom setting; (4) student learning experiences; and (5) student learning outcomes. Further research is needed in classroom learning, classroom teaching, teacher competence, and the relationship between pre-existing teacher characteristics used for admission to teacher education schools and later competence. The validity of measures of teacher competence depends on the content validity of the measures of student learning outcomes used, and how well the outcomes attributable to the teacher are separated from the outcomes that are due to context factors. The expected gain score is useful in estimating the post-test score that a student with a specified pre-test score would achieve. A review of literature on teacher competencies led to the development and validation of a measure for Virginia's competency-based certification system. Seventy behavioral indicators of fourten teacher competencies resulted. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED261085

ISSUES AND PROBLEMS IN THE VALIDATION OF TEACHING AND TEACHER PROFESSIONAL BEHAVIORS

by

Donald M. Medley

University of Virginia

an invited address to Division D

American Educational Research Association.

Chicago, Illinois

April 3, 1985

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Medley, Donald M.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 505

The topic we have been asked to discuss today, **Issues and Problems in the Validation of Teaching and Teacher Professional Behaviors,** is a very large one, and its boundaries are sufficiently vague to give us considerable freedom in deciding exactly what we will talk about. I plan to comment briefly on validation problems and issues as they relate to the kind of teacher evaluation systems that have been springing up all over the country mainly as the result of mandates issued by state legislators who seem to be unaware that such issues and problems exist.

I will begin by presenting my own model of the process of teacher evaluation, and will organize the rest of my comments around it. I will also introduce a statistical procedure for assessing teacher effectiveness that nobody else seems to be using as yet, and illustrate its application to some evaluation problems.

I am making one important assumption. I am assuming that whenever and however we evaluate teachers we do so to find out how effective they are, how much the average student may be expected to learn from each of them; and a valid system is one that gives us this information.

There are five major targets on which evaluations of teachers or their teaching may be focussed for different purposes.

1. Teacher evaluation may focus on personal characteristics of candidates for admission into pre-service professional teacher preparation programs.

2. Teacher evaluation may focus on knowledges, skills, attitudes, etc., that teachers possess at the end of a period of professional training.

3. Teacher evaluation may focus on the actual behaviors of teachers while teaching.

4. It may focus on the behavior of a teacher's students while they are being taught.

5. Finally, teacher evaluation might focus on the knowledge, skills, attitudes, etc. that the teacher's students possess after the teaching stops.

I will refer to these five foci as **assessment points** and number them one to five, respectively.

Different characteristics of teachers and different aspects of

teaching behavior are accessible for evaluation at each of these five points, and these differences are important. In order to help us keep them clear in the discussion to follow, I have assigned names to the characteristics assessed at each point. The names of the characteristics assessable at each point and the relationships among them are shown schematically in Figure 1.

(Insert Figure 1)

Pre-existing teacher characteristics are assessed directly at Point 1. They include characteristics of two kinds, first, knowledge, skills, attitudes, etc., that the candidate will need in order to do satisfactory work in the teacher preparation program; and, second, knowledge, skills, attitudes, etc. the candidate will need in order to succeed as a teacher, but will not acquire as part of her professional training. [1]

Teacher competence is assessed directly at Point 2. A competent teacher is one who possesses a minimum set of teacher competencies. A **teacher competency** is a unit of knowledge, a skill, or a set of attitudes required for satisfactory performance as a teacher.

Teacher performance is assessed directly at Point 3. Teacher performance refers to the pattern of behavior a teacher displays while teaching a class.

Teacher competence is a personal characteristic of the teacher that the teacher carries with her from one setting to another, but **teacher performance** is a function of setting and will change from one setting to another.

Student learning experiences are assessed directly at Point 4. Student learning experiences are activities of students that result in learning.

Student learning outcomes are assessed directly at Point 5. Student learning outcomes are changes in student knowledge, skills, attitudes, etc. measured after teaching ends. The related term **teacher effectiveness** refers to the portion of these outcomes that is attributable to the efforts of the teacher.]

The five assessment points in the diagram are joined by arrows to indicate the relationships among the different aspects of teaching assessed. Notice that each aspect is directly influenced by the one at the point immediately preceding it, that is, the one to its left in the diagram. Notice also that each aspect directly influences the one next to it on the right. Each

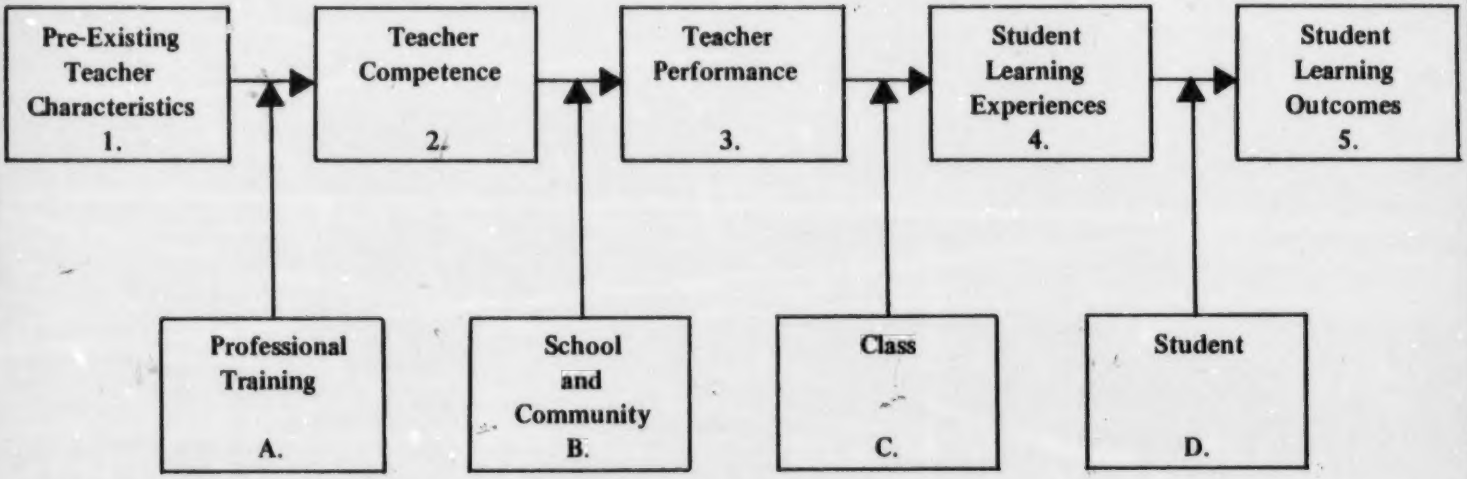


Figure 1. Five Assessment Points in a Teacher's Career and Four Sources of Context Factors

characteristic or aspect measured is affected directly or indirectly be every aspect to its left, and affects all aspects to its right directly or indirectly.

Student learning outcomes, for instance, are directly affected by student learning experiences, but are affected only indirectly by teacher performance, and even less directly by teacher competence. This reflects the fact that learning is something each student must do for himself, and that the role of the teacher in the educational enterprise is to see that students have learning experiences likely to result in the kinds of learning outcomes schools are supposed to produce.

The terms **teacher competence, teacher competency, teacher performance,** and **teacher effectiveness** are often used interchangeably in discussions about teacher evaluation. This is unfortunate, because unless clear distinctions are somehow made among these four concepts, it is difficult if not impossible to communicate some very important ideas. Failures to make these distinctions account for much of the confusion that plagues discussions of teacher evaluation. It is to avoid such ambiguities that I have assigned the four terms to the four concepts. And although there is an element of arbitrariness in the way in which I have done so, the usage I propose corresponds fairly well with common usage.

The diagram in Figure 1 also shows variables of four other types that it is important for us to consider when we attempt to validate measures of teachers and teaching. These factors, often called **context factors**, are not under the control of the teacher, but their potential influence on assessments made at points 2 to 5 is too strong to be ignored. Each type of context factor is labeled according to its principal source, and arrows indicate the assessment point directly influenced by each of them.

Type A factors include characteristics of professional training which directly influence teacher competence. Type B factors include characteristics of the school and community in which the teaching takes place which directly influence teacher performance. Type C factors include characteristics of the teacher's class which directly influence the learning experiences students have in that class. And Type D factors include characteristics of individual students that directly influence learning outcomes.

Our assumption that the purpose of any teacher evaluation made at any assessment point is to discriminate more effective teachers from less effective ones has one apparent implication that needs clarification. This assumption seems to imply that the only criterion against which any evaluation system should be

validated is a measure of student learning outcomes made at point 5. Although incorrect, this conclusion is logical, and has great intuitive appeal to most of us.

What could be a more reasonable policy than one which admits into a teacher preparation program only those candidates who can ultimately be expected to become effective teachers? Such a policy would seem to require any admissions test to be validated by correlating it with a measure of how effective candidates become after they finish training and become practicing teachers.

And what could be more logical than to require every teacher to acquire and demonstrate whatever set of competencies is necessary to be effective in the classroom before she is certified competent to teach? This would seem to require that each competency in the set be significantly correlated with a measure of how effective a teacher with that competency becomes.

Logical though these ideas may be, a glance at the diagram indicates how impractical they are. Five different types of variables intervene between a measure of teacher competence made at Point 2 and a measure of learning outcomes made at Point 5, and seven types of variables intervene between a measure of pre-existing teacher characteristics made at Point 1 and a measure of outcomes made at Point 5. Each of these variables attenuates any relationship that exists between the measures to be validated and the criterion used. It is rarely practical or even possible to measure and make statistical adjustments for enough of these variables so that validity can be demonstrated when it is present.

The practical solution to the validation problem is to solve it a step at a time, that is, by correlating a measure obtained at any point (except Point 5, of course) with one made at the next point on its right in the diagram. Measures of competence made at Point 2 would be correlated with measures of performance at Point 3 instead of with measures of effectiveness made at Point 5. Since the measure to be validated would then be one that directly influenced the criterion measure used, the relationship should be relatively strong and attenuated by context factors of one kind only.

This approach is as logical as the other. It is perfectly reasonable to argue that a teacher preparation program should admit only those candidates who will be competent teachers when they finish the program, and therefore to require that any admissions test used should predict scores on a measure of competency administered at the end of the program. Similarly, it can be argued that only teachers who may be expected to perform

satisfactorily on the job at Point 3 should be certified as competent at Point 2. A competency measure may therefore be considered valid if scores on it predict scores on a measure of performance administered after the teacher has obtained a full-time teaching position.

Unfortunately, in the present state of the art of teacher evaluation, this second approach is not very practical either. Successful implementation of this approach would require that previously validated evaluation procedures be available at each assessment point (except Point 1, of course), before measures made at any earlier point could be validated. But at this point adequately validated measures are mot available at any of the four points, much less at all of them.

Perhaps we should view the diagram in Figure 1, not as an evaluation model, but as an agenda for research in the teaching-learning process.[2] Seen in this light, it indicates a need for four kinds of research.

We need, first of all, **research in classroom learning,** research into the relationships between student learning experiences in the classroom measured at Point 4 and student learning outcomes measured at Point 5. We need to know how students' time in the classroom can be most profitably utilized. The concept of "academic learning time" is an example of the kind of thing this research should produce.[3]

Next we need **research in classroom teaching,** research relating teacher performance measured at Point 3 to student learning experiences measured at Point 4. Some of the most productive research in teaching has been process-product research, [4] which has sought relationships between measurements of teacher performance made at Point 3 and measures of student learning outcomes made at Point 5. These relationships are attenuated by two kinds of context factors (C and D), as well as by differences in student learning experiences. When valid measures of student learning experiences become available process-product research should become obsolete.

Third, we need **research in teacher competence,** research into the relationships between the competencies a teacher possesses and her performance in the classroom. The fact that past efforts to relate teachers' scores on teacher competency tests to measures of student learning outcomes have not been successful [5] suggest that we should change our strategy, and try to relate teacher competency test scores to measures of teacher performance instead.

Finally, there is a need for **admissions research,** research into

the relationships between pre-existing teacher characteristics measured at Point 1 and teacher competence. measured at Point 2

Research of all of these types is relatively costly; and the principal cost is the cost of acquiring data in the quantities needed. It will not have escaped your notice that the relationships this research is supposed to establish are identical with those needed to provide the evidence of validity needed by the school systems and state education agencies that are developing these evaluation systems and administering them to thousands of teachers. The data needed for the research studies and the data needed for the validation studies are the same data, large amounts of which are already being collected in the programs that need the validity evidence. If --and when-- the researchers and the evaluators begin to collaborate we may hope to see rapid progress on both sides. I think I see signs that this beginning to happen.

Let me turn now from these generalities to some more specific suggestions growing out of my own encounters with members of both groups.

Notes on Validating Direct Measures of Teacher Effectiveness.

Since we have assumed that evaluating teacher effectiveness, a teacher's ability to promote student learning, is the ultimate objective of any effort to evaluate teaching, one might wonder why so much teacher evaluation is based on measures of teacher performance (Point 3) or of teacher competence (Point 2) instead of on direct assessments of teacher effectiveness made at Point 5.

There are at least three cogent reasons for this. One is that measures made at Point 5 do not contain any diagnostic information. It is not much use to a teacher to be told that she is ineffective. She probably knows that already. What she needs to know is what to do in order to become effective; what competencies she needs to acquire or how to make better use of the ones she has in her teaching behavior. Evaluations of teacher competence or teacher performance are much more useful in most cases than direct evaluations of teacher effectiveness.

A second reason why direct measures of teacher effectiveness are not used more frequently is that the results are not available soon enough to be of any use. It takes time, usually several months, before most of the important changes teachers try to produce can be detected and reliably measured, and most of the decisions that should be based on the evaluations cannot wait that long. Information about teacher performance or competence is available when it is needed.

But there is a third reason for not using direct assessments of teacher effectiveness that is more important than either of these: the fact that they are too unreliable to be useful. Research indicates that the average correlation between two measures of the effectiveness of the same teacher obtained in different classes is about .30. This means that an elementary teacher who is the most effective in a school during one year is almost as likely to be the least effective as she is to be the most effective during the following year. [6] This is a poor basis for deciding which teacher should get tenure or which should be given merit pay. The instability of direct measures of teacher effectiveness is sufficient reason all by itself for not using them to evaluate individual teachers.

The appropriate use to be made of direct measures of teacher effectiveness as criteria against which to test the validity of other, more timely, more practicable and more reliable teacher evaluation procedures? Low reliability does not preclude such uses if the measures are valid.

The validity of a direct measure of teacher effectiveness depends on two things. One is the content validity of the measures of student learning outcomes used. The other is how well the portion of the outcomes that may be attributed to the efforts of the teacher is separated from the portion that is not attributable to her efforts, but to context factors.

One major context factor is that measured by an appropriate pretest. Unless students are randomly signed to classes taught by different teachers, there will be differences between mean post-test scores of students in different classes that exceed those expected by chance; and these will result in non-chance differences in mean post-test scores. Teachers whose students' pretest scores happen to be above average will therefore appear to be more effective than they are, and teachers whose students' pretest scores happen to be below average will appear to be less effective than they are.

A number of different statistical techniques have been used to compensate for this. The more respectable ones use the regression of post-test scores on pre-test scores to "adjust" the mean post-test scores to what they would have been if all classes in the sample had had the same mean pre-test score or scores. The adjusted mean post-test scores are then used as measures of the teachers' relative effectiveness.

The main difference between the different techniques that have been used for this purpose in the past has been in whether they make use of the information contained in the correlations between

pretest and post-test scores within classes; and if they have used it, how they have used it. Those that have used this information have assumed that these correlations are the same in different classes, and have therefore ignored the information contained in differences in these correlations from class to class. These differences are manifested in differences in the slopes of the regression lines within different classes.

The slope of the regression line within a class indicates whether (and to what extent) the amount any individual student learns in that class depends on his ability (or whatever characteristic the pretest measures). A steep slope indicates that more able students learn more than less able ones do in that class. A flat slope indicate that students of all levels of ability learn approximately the same amount. Much more sophisticated discussions of this problem exist; the important point is that that these slopes contain essential information about the effectiveness of the teachers. [7]

Expected Gain Scores. I would like now to describe a simple technique for making use of this information that I have found extremely useful. Instead of the usual adjusted mean gain score it yields a statistic which I shall call the **Expected Gain Score.** The expected gain score in a class estimates **the post-test score that a student with a specified pre-test score would get in that class.**

Consider, for example, the expected gain score of a student whose pre-test score equals the mean of the local norms for his grade and subject. This expected gain score estimates the post-test score that a student of average ability would get in a class. (This should *not* be confused with the post-test score that the average student in that class would get.) Variations in this score between classes reflect differences in the effects of the learning experiences the average child in that grade and school would have in the different classes.

The expected gain score in each class is estimated from the regression of post-test scores on pretest scores **within that class only.** Figure 2 shows three (hypothetical) regression lines estimated in three fourth-grade classes marked A, B, and C. The pre-test score of the average student in this grade and school is 34, and a vertical line has been drawn to indicate the point on each regression line that corresponds to that pretest score. In each case, this point is the predicted post-test score of any student with that pretest score in that class, which we have decided to call the *expected gain score* for the average student in that class. The approximate values of the three expected gain scores are 44 in class A, 45 in class B, and 49 in class C. This indicates that the average student would learn more in class C

than he would learn in either of the other two classes.

(insert Figure 2)

The expected gain score is not a measure of teacher effectiveness; it is a measure of student learning outcomes in a teacher's class that holds constant whatever student characteristics (Type D context factors) the pretest measures. To repeat, variations in the expected gain score of the average student in different classes clearly reflect differences in the learning experiences average students have in different classes.[8]

Part of these variations may may be due to differences between the classes rather than differences between the teachers. The kind of experiences a student has in a class and what he learns there may depend on class characteristics like average ability, attitude toward school, or group cohesion. These are examples of Type C context factors which must be controlled before measures of teacher effectiveness can be derived from expected gain scores. Perhaps the simplest way of doing this is to partial out measures of Type C context factors from correlations between expected gain scores and measures of characteristics assessed at Point 3 or before. When expected gain scores are correlated with teacher performance measures, for example, partial correlations holding Type C factors should be used.

Whether or not a teacher performance measure that correlates with expected gain scores is really valid will then depend on the kind of student learning outcomes the post-test measures, that is, on the content validity of the post-test scores as measures of what the students are supposed to learn in school. As we shall see, there is evidence that what kind of teacher performance correlates with student learning outcomes depends on what kind of student learning outcomes we measure.

I would like to use an experiment in classroom learning reported by Wagner. [9] to illustrate this point, and at he same time to serve as an example of one way of validating a measure of student learning experiences.

Notes on Validating Measures of Student Learning Experiences.

Wagner's experiment was conducted in six elementary school science classes. All six classes studied the same content; but in three classes (called "inquiry classes") the students performed simple experiments and discussed the results in class while in the other three classes (called "reading-recitation"

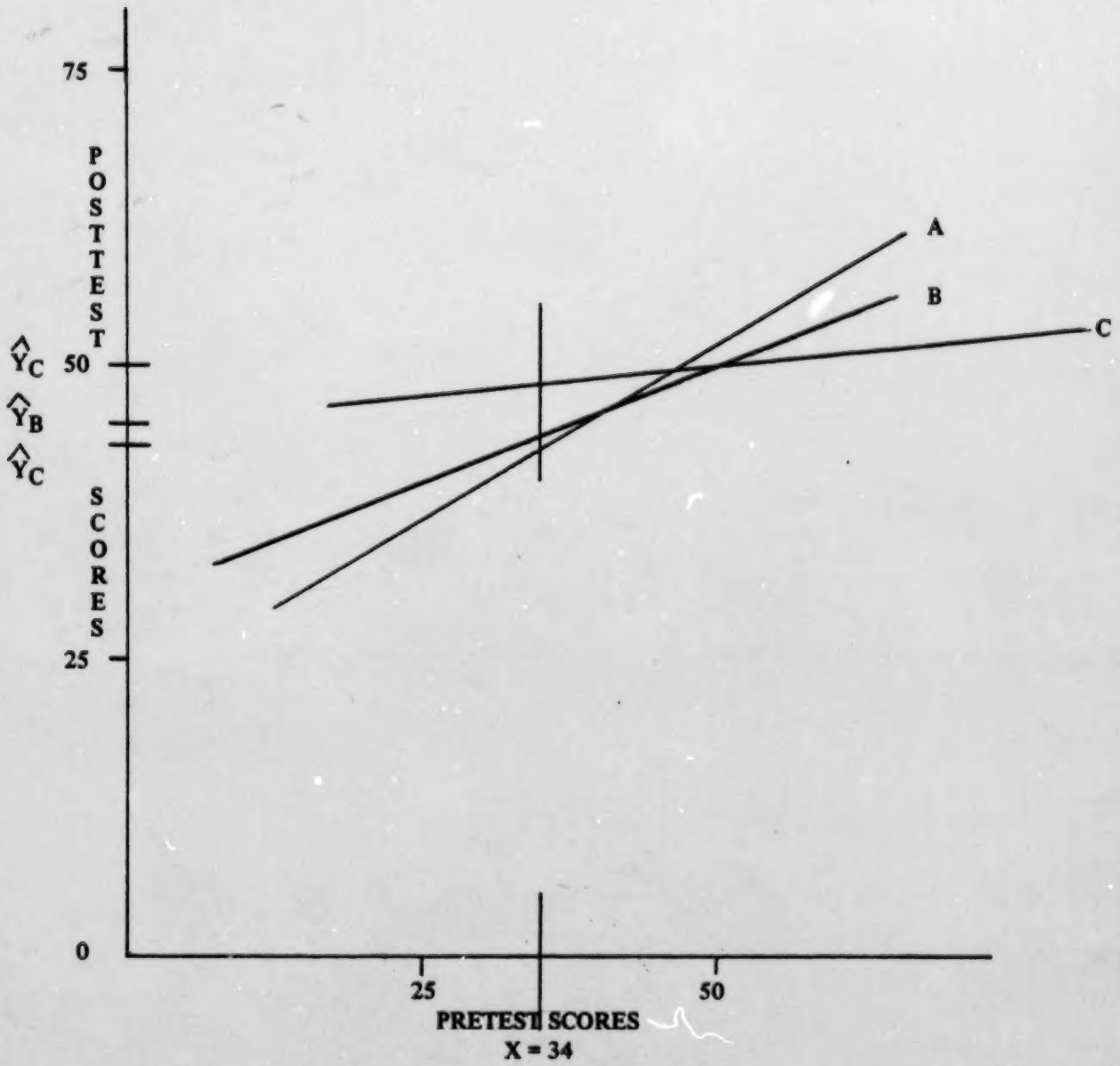


Figure 2. Within-Class Regression Lines for 3 Hypothetical Classes

classes) the students did no experiments. Instead, they read descriptions of experiments which they discussed in class.

The same four tests were administered pre and post in all six classrooms. Two of the tests, called "content" tests, were teacher-made tests of knowledge of the factual content taught; the other two, called "process" tests, were standardized tests designed to measure understanding of the scientific method independent of content. Expected gain scores were calculated on each test, and expressed in standard or "z" score form.

A measure of the quality of student learning experiences in each class was obtained from observations recorded on OScAR 5V and scored on a key related to the quality of the discussions observed. [10] If we think of this experiment as one intended to validate this measure of student learning experiences (Point 4 in Figure 1), differences between expected gain scores in the two types of classes may be be regarded as evidence of its validity.

Figure 3 shows the mean expected gain scores on the two kinds of test in each type of class. The results clearly indicate that the average student in this school would learn a great deal less about the scientific method in a reading-recitation class than he would learn in an inquiry class; but they not indicate that he would learn any more of the factual content being taught in one type of class than in the other.

(Insert Figure 3)

Of course the experiences these students had in the two types of classes differed in other ways besides discussion quality, in particular in whether they did any experiments. But when the two mean expected gain scores were adjusted by analysis of covariance to remove the measured differences in discussion quality, the differences in expected gain scores on the process tests were reduced by about half. This was taken as evidence that the differences in learning outcomes on the process measures were at least partially related to discussion quality, It was also interpreted as evidence of the validity of our measure of student learning experiences as a predictor of learning outcomes, when those outcomes were measured on a test of understanding of the scientific method.

The substantive findings of this experiment, sensible though they are, should be viewed as highly tentative; but they do illustrate two points. One is the importance of using the the right type of student learning outcomes as a criterion in a validation study. The other is that an old-fashioned methods

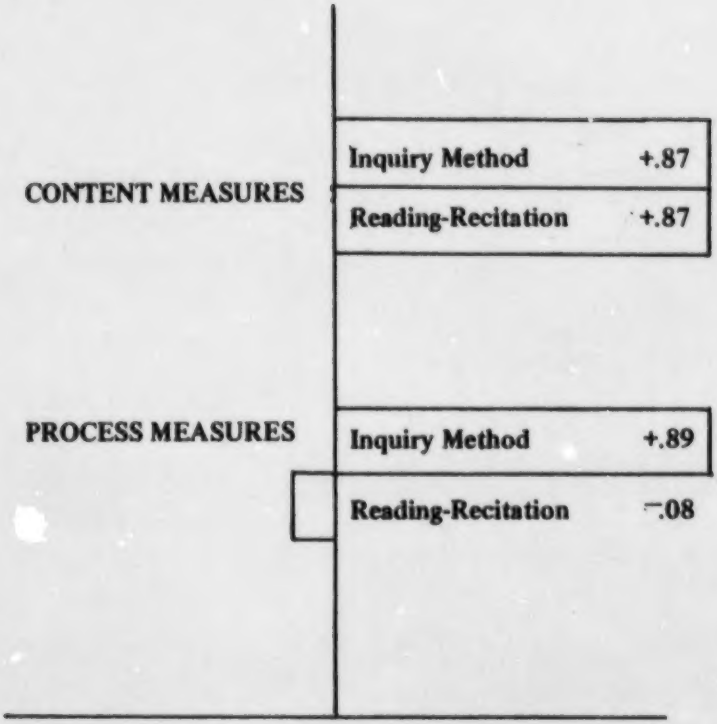


Figure 3. Achievement Gains on Two Types of Measures in Junior High School Science Classes Taught by Two Different Methods

experiment is not a bad design for validating measures of student learning experiences if a measure of the learning experiences (Point 4) is obtained along with the measure of learning outcomes (Point 5).

Notes on Validating Teacher Performance Measures.

As our knowledge of classroom learning increases, as we develop more and more valid measures of student learning experiences related to desired learning outcomes, research in classroom teaching will become both more feasible and more important. It will become easier to establish which teaching behaviors are most effective in providing students with those learning experiences that are most closely related to specific learning outcomes.

We have noted how useful process-product research has been in relating measures of teacher performance to measures of student learning outcomes. It has succeeded despite the fact that the relationships it seeks are indirect, that is, ones that involve measurements made at non-adjacent assessment points.

Process-product research should continue to play an important role in research on teaching for some time to come. But, because the relationships it seeks are indirect, it is important to take full advantage of anything that can be done to make the relationships easier to detect. One way to do this is by getting better control of Type C context factors -- those associated with the class the teacher teaches.

Expected gain scores are particularly useful here because they make it possible to measure and control Type C and Type D context factors separately. An example of this is provided by a process-product study reported by Lara [11]

Lara's study involved 21 elementary school classes, 15 in reading and six in arithmetic. Parallel forms of an appropriate subtest of the Iowa Tests of Basic Skills were administered in each class in the fall and the spring. Each teacher's classroom behavior was observed on several different occasions during the intervening school year and recorded on one of three observation schedules. The records were scored on 22 different dimensions of teacher performance, each one of which had been identified by a panel of classroom teachers as one that should correlate positively with student learning outcomes, that is, as a "behavioral indicator of competence."

The use of expected gain scores made it possible to test an assumption implicit in most process-product studies: the assumption that the most effective pattern of teacher performance is the same for students at all ability levels. To test this

assumption, Lara computed not one but two expected gain scores for each of the 21 classes, one for high-ability students and one for low-ability students. A high-ability student was defined as one whose pre-test score was at the 86th percentile of the local norms for the grade he was in; and a low-ability student was defined as one whose pre-test score was at the 16th percentile of the same group.

Two Type C context factors were assessed in the study. There is a widespread popular belief that a low-ability student will learn more in a class in which most of the students are more able than he is than he would learn in a class where most students' ability is equal to or below his. There is a companion notion that a high-ability student will learn less in a class in which most students are less able than he than he would in a class in which most students' ability equals or exceeds his own. If either of these beliefs were correct, differences in class means would attenuate the process-product correlations obtained. The mean pretest score in each class was used to estimate the average ability of that class.

The second Type C context factor assessed was the heterogeneity of ability in each class, measured by the standard deviation of pretest scores in it. One of the oldest debates in education has to do with whether the heterogeneity or homogeneity of ability in a class has anything to do with how much the students in a class learn. If it does, then differences in standard deviations of different classes would attenuate the process-product correlations obtained.

If either of these two context measures was significantly correlated with expected gain scores at either level, controlling it by partialing it out of the process-product correlations should increase the sensitivity of the analysis, that is, make it easier to detect weak relationships. As it turned out, neither the average ability nor the heterogeneity of ability of a class correlated with expected gain scores at either level, so zero-order correlations were used to "validate" the 22 behavioral indicators of competence.

Correlations were calculated between scores on each of the 22 "performance competencies" and each of the two expected gain scores separately for the two subjects. The 88 correlation coefficients obtained are shown in Figures 4 and 5. The 22 pairs of correlations obtained for each competency indicator in the 15 reading classes are shown in Figure 4. One profile represents the behavior pattern most effective with high-ability students, the other represents the pattern most effective with low-ability students. Although the two profiles are somewhat similar in shape, they are by no means identical.

(Insert Figures 4 and 5)

Correlations obtained in the six arithmetic classes are shown in Figure 5. As might be expected with such a small sample, these correlations show more variability and the profiles differ more. Despite the small sample size, The two correlations with the same behavior differed significantly in nine of the 22 pairs. The teachers most effective with high ability students behaved very differently than those most effective with low ability students.

Once again I am presenting these findings, not for their substantive interest, but as illustrations of some of the possibilities opened up by the use of expected gain scores instead of adjusted mean gains or similar statistics.

Notes on Validating a Measure of Teacher Competence

What I plan to say about validating measures at Point 2 will be based on what I have learned as a member of a team at the University of Virginia which is responsible for developing a competency-based teacher certification system for the state of Virginia. The contract specifies, among other things, that each beginning teacher must demonstrate a minimum set of competencies in her behavior with her own class, and that each of the competencies she is required to demonstrate must be consistent with the findings of research in teacher effectiveness. The first of these conditions requires us to build a system for inferring competence (Point 2) from performance (Point 3) rather than a test that purports to measure competence directly. The second requires us to present evidence that the competencies in the set we measure are also part of the set of competencies that research has found to be related to teacher effectiveness.

Even if this second condition did not require us to establish the validity of the evaluation system on the basis of content validity, it would make sense to do so because neither of the empirical strategies described earlier in this discussion is feasible. We have seen how difficult it is to detect relationships between measures of competence (Point 2) and measures of student learning outcomes (Point 5). And we cannot use the alternate strategy and correlate competence with performance (Point 3) because our measures of competence are actually made at Point 3.

The process of content validation begins with the definition of the domain to be measured by the evaluation system. Since teaching is a learned profession, this domain should be a domain

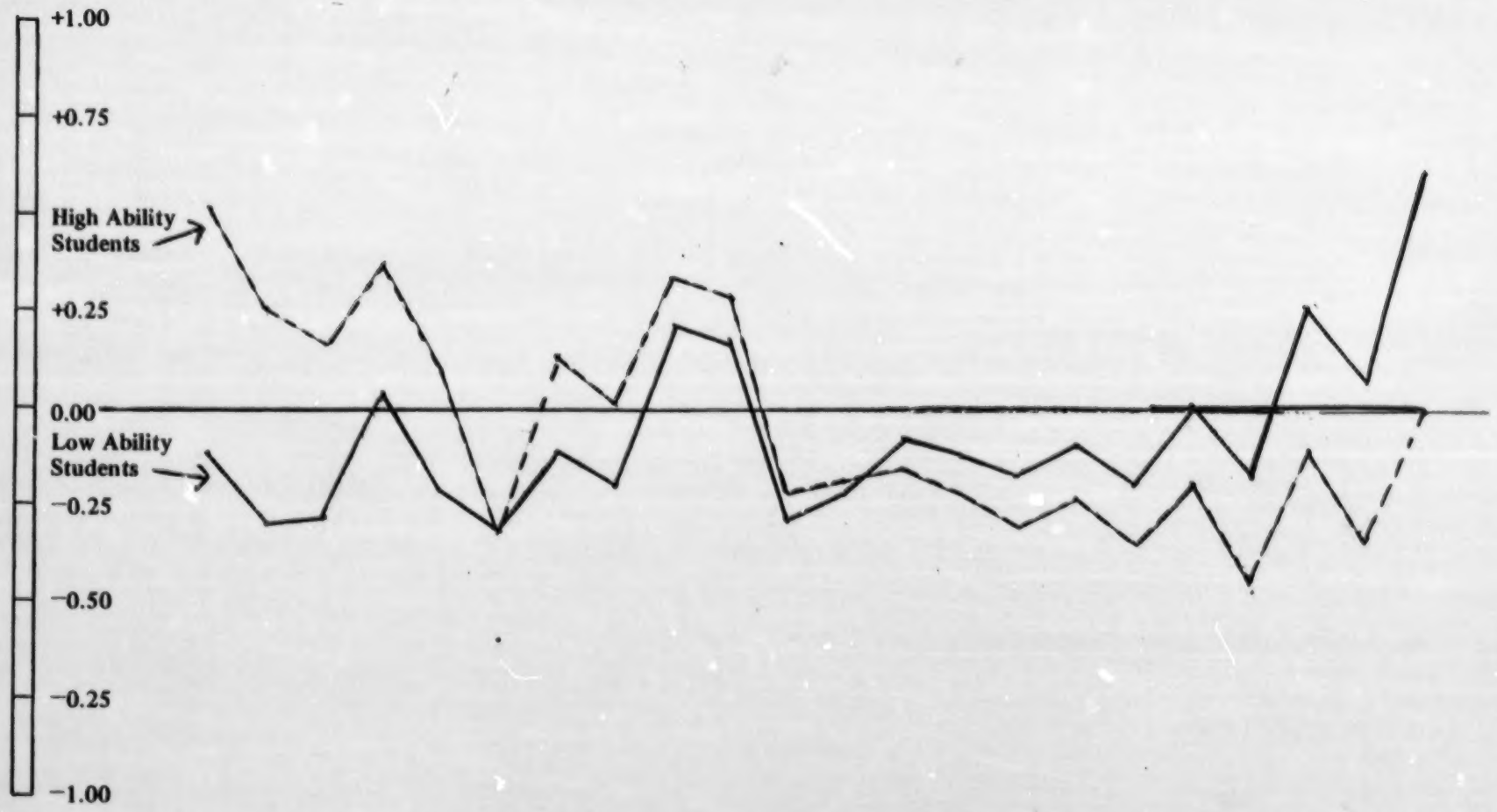
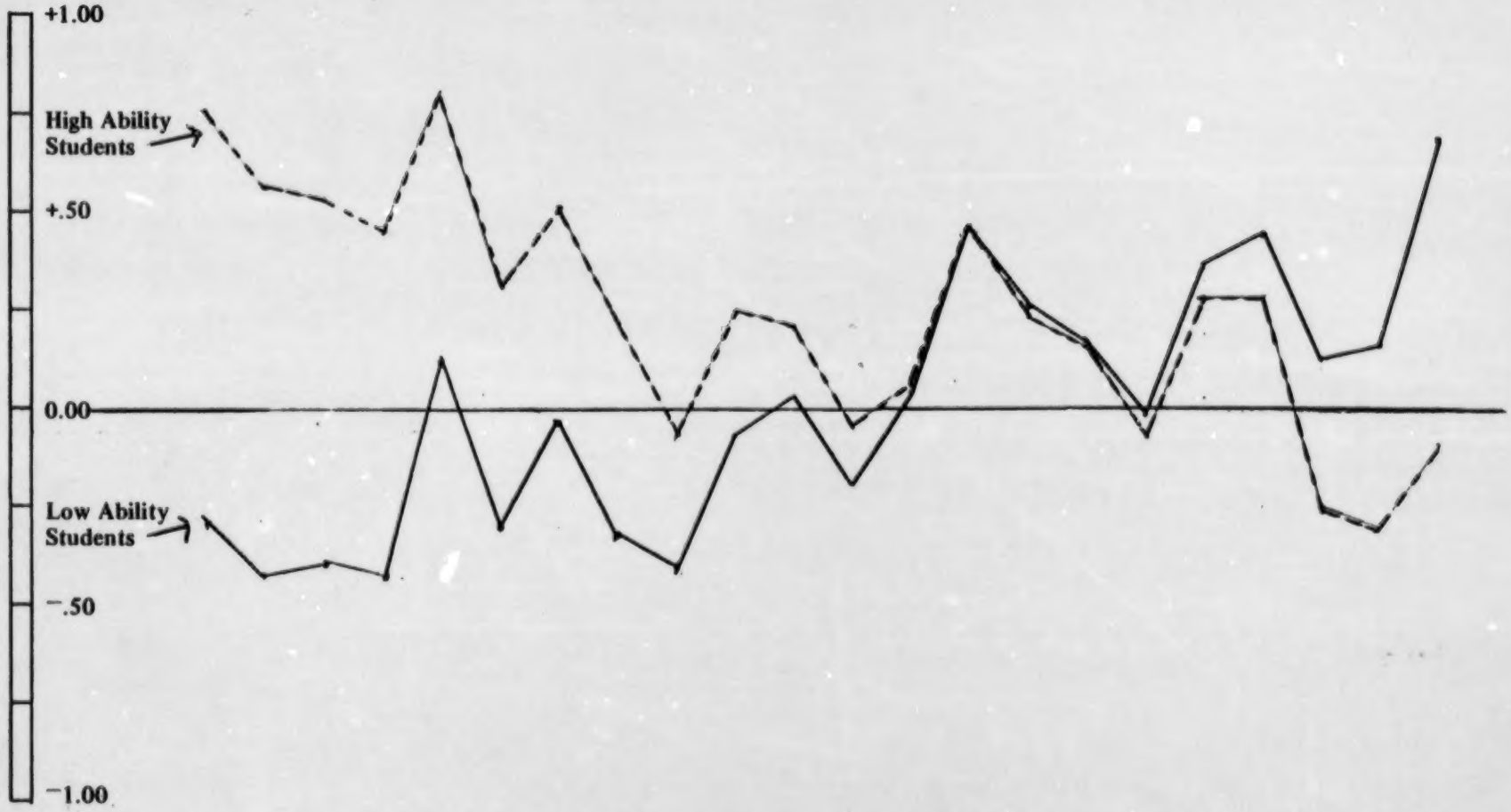


Figure 4. Correlations Between Scores on 22 Teacher Behavior Measures and Expected Gain Scores of High and Low Ability Students in 15 Reading Classes



-136-

Figure 5. Correlations Between Scores on 22 Teacher Behavior Measures and Expected Gain Scores of High and Low Ability Students in 6 Arithmetic Classes

of professional knowledge; and in order to be demonstrable in performance, the knowledge must be functional, that is, the teacher must have access to it while she is teaching, so that it must be accompanied by whatever skills are needed to implement it. The knowledge must have a base in the empirical research. And in addition it must be acceptable to the Virginia educators and teacher educators who must eventually use it or teach it to others.

We began our efforts to identify the competencies that the domain should include by collecting and examining lists of teacher competencies compiled by others. We found to our surprise that instead of listing **competencies** a teacher needs in order to perform the functions of a teacher these lists specified the **functions** a teacher should be able to perform. They contained dimensions of teacher performance, such as **skill in classroom management** and **ability to teach concepts.** But we were looking for lists of concepts, for such things as **academic learning time** and **reinforcement theory**, knowledge of which is expected to contribute to the successful performance of not just one but many teaching functions. This is the knowledge a teacher must acquire in her preservice training rather than through experience. These so-called competency lists seemed to us to define **teacher performance** rather than of **teacher competence.**

Visualize a two-way grid with teaching functions on one axis and teacher competencies on the other. The cells contain specific teaching behaviors called **behavioral indicators of competence.** The behaviors in each cell indicate possession of a competency listed on one axis and contribute to the performance of a function listed on the other. Since our charge is to certify teachers on the base of competence, not performance, we will require a beginning teacher to demonstrate possession of the competencies, but not the ability to perform the functions.

The domain of competencies is specified at two levels. At one, it can be described in terms of fourteen **competencies**, fourteen areas of functional professional knowledge at one level. At another level it can be described in terms of 70 behavioral indicators of competence. Each of the fourteen competencies is operationally defined in terms of two or more behavioral indicators. The presence of any one of these indicators in a teacher's behavior may be taken as evidence (positive or negative) that the teacher possesses one of the fourteen competencies as defined and therefore as evidence that she is competent (or incompetent) to teach.

The 70 behavioral indicators were originally identified through a search of recent reviews of the research in teacher effectiveness. [12] This is prima faciae evidence of content

validity, in the sense that the domain of competence measured is consistent with the research. This will be verified by having the 70 indicators reviewed by a panel of researchers expert in this literature.

Evidence that the fourteen competencies in the domain are acceptable to Virginia educators is based on their participation in the process by which the competencies were identified and defined. An advisory board appointed by the state board and representative of educators and teacher educators in the state worked closely with the project team every step of the way.

The claim we make is that any teacher who performs at a level yet to be determined on this system will have demonstrated possession of a set of valid functional professional knowledges sufficient to justify certifying her as competent to enter the practice of teaching in the state of Virginia. She will be competent enough, will know enough about teaching and learning to become an effective teacher with experience. If she does not do so, the failure will not be due to a lack of minimal competence on entry.

Concluding Remarks.

I hope that most of you will remember two things I have said today. I hope you will remember something about the model of teacher evaluation I have described and the five assessment points it contains. I hope you will remember that the characteristics of teachers that are available for evaluation at different points are very different; that they are measured by different methods and used for different purposes, and especially that they are validated in different ways.

I hope that you will also remember the statistic I have called the expected gain score, which, because it estimates how much the same student would learn in different classes on the basis of how much students learn in those classes, provides a sounder and simpler basis for evaluating teacher effectiveness than any of the procedures in common use today.

NOTES

1. In order to avoid pronoun tangles, I will consistently use feminine pronouns to refer to teachers and masculine pronouns to refer to students or others.

2. Cf. Medley, D.M. Research on teaching, evolution of. **International Encyclopedia of Education,** T. Husen and T.N. Postlethwaite, editors. Oxford, Pergamon Press, 1985.

3. Berliner, D. L. *Tempus educare.* **Research on Teaching: Concepts, Findings and Implications.** P.L. Peterson and H.J. Walberg, editors. Berkeley, CA, McCutchan, 1979.

4. See, for example, **Domains: Knowledge Base of the Florida Performance Measurement System.** Chipley, Florida, Panhandle Area Educational Cooperative, 1983; Brophy, J.E. Teacher behavior and its effects, **Journal of Educational Psychology,** 1979. **71:** 733-750; Good, T.L. Teacher effectiveness in the elementary school. **Journal of Teacher Education,** 1979 **30, No. 2:** 52-64; Medley, D.M. *Teacher Competence and Teacher Effectiveness: A Review of Process-Product Research* Washington, D.C., American Association of Colleges for Teacher Education, 1977; Rosenshine, B. **Teaching Behaviors and Student Achievement.** London, National Foundation for Educational Research, 1971.

5. Cf Barr, A.S. The Measurement and prediction of teaching efficiency: a summary of investigations. **Journal of Experimental Education,** 1948. **16:** 203-283; Quirk, T.J., B.J. Witten, and S.F. Weinberg. Review of studies of the concurrent and predictive validity of the National Teacher Examinations, **Review of Educational Research,** 1973, **43:** 89-113.

6. See Rosenshine, B. The Stability of teacher effects upon student achievement. **Review of Educational Research,** 1970, **40:** 647-652; Brophy, J.E. Stability in teacher effectiveness. **American Educational Research Journal,** 1970, **10:** 245-252.

7. See, for example, Burstein, L. and R.L. Linn, Analyzing multilevel data in the presence of heterogeneous within-class regressions. **Journal of Educational Statistics.** 1978, **3:** 347-383.

8. More than one pretest can, and in most cases probably should, be used; but for the sake of simplicity, only one will be used in the examples given.

9. Wagner, A.M. **A Study of Elementary School Science Achievement Related to Teacher Questioning Styles and Methods of Instruction.** Unpublished doctoral dissertation, University of Virginia, Charlottesville, Virginia, 1983.

10. Medley, D.M. **The Development and Use of Observation**

Schedule and Record, Form 5V. Unpublished manuscript, Curry School of Education, University of Virginia, Charlottesville, VA, 1985.

11. Lara, A.V. **[Pupil Ability as a Moderator of Correlations between Teacher Behavior Patterns and Pupil Gains in Reading and Mathematics.** Unpublished doctoral dissertation, University of Virginia, Charlottesville, Virginia, 1983.

12. SeeNote 2, above